Automated Generation of Cognitive Ontology via Web Text-Mining

Richard Gao (rigao@ucsd.edu)
Thomas Donoghue (tdonoghue@ucsd.edu)
Bradley Voytek (bvoytek@ucsd.edu)

Department of Cognitive Science, UC San Diego
9500 Gilman Drive, La Jolla, CA 92093-0515, USA

Abstract

A key goal of cognitive science is to understand and map the relationship between cognitive processes. Previous works have manually curated cognitive terms and relations, effectively creating an ontology, but do they reflect how cognitive scientists study cognition in practice? In addition, cognitive science should provide theories that inform experimentalists in neuroscience studying implementations of cognition in the brain. But do neuroscientists and cognitive scientists study the same things? We set out to answer these questions in a data-driven way by text-mining and automated clustering to build a cognitive ontology from existing literature. We find automatically generated relationships to be missing in existing ontologies, and that cognitive science does not always inform neuroscience. Thus, our work serves as an efficient hypothesis-generating mechanism, inferring relationships between cognitive processes that can be manually refined by experts. Furthermore, our results highlight the gap between theories of cognition and the study of their implementation.

Keywords: ontology; cognitive neuroscience**;** text-mining; neuroinformatics, meta-analysis

Introduction

Ontology: Key Challenge in Cognitive Science

One of the fundamental goals of cognitive science is to study the set of processes that combine to give rise to "cognition". These processes can be thought of as abstractions to common, overlapping sets of behaviors. Constrained by methodological behaviorism, we can only observe behavior and label underlying cognitive processes after the fact. As such, they do not have direct grounding in the physical world, and thus need to be defined by the relational structure that link each other – an ontology. For example, attention and working memory are processes with different labels but are nonetheless woven together through behavior: one cannot allocate "working memory" without "paying attention". Thus, as we collect more observations to fill up the space of cognitive processes, we must be attentive in organizing what we know. This is the problem of mapping the ontological structure of cognitive processes, and has received extensive consideration previously (see, e.g., Poldrack & Yarkoni, 2016).

Neuroscience: Studying the Substrate of Cognition?

If cognitive processes are viewed as algorithms performing a set of computations, there must then exist a physical substrate that is performing the computations. In the case of computer algorithms, the substrate consists of transistor elements. The brain, on the other hand, is a large part of the computational substrate of human cognition (along with our body and environment). Cognitive neuroscience, with the aid of neuroimaging, has revealed much about our cognitive processes, such as timing between consecutive steps in a cascade of processes. However, neuroimaging studies are almost always conducted in the laboratory, with specific physical and task constraints. Hence, one cannot be certain that cognitive neuroscience actually measures, or even attempts to measure, the full array of cognitive processes at play. For example, the consolidation of long-term memory is quite difficult to measure within the span of a single experiment, while visual perception can be easily studied. Conversely, observations in neuroscience may provide constraints for cognitive theories, but only if there is an overlap of interest in the same processes. Thus, we should understand the degree to which we are over- and under sampling cognitive processes while measuring the brain. This is the problem of adequate sampling of cognitive processes in cognitive neuroscience.

Frameworks for Ontology-Mapping

The problem of ontology has been addressed previously. Notably, Poldrack and colleagues (2011) started a monumental effort in charting the ontological space of cognitive processes, as well as their related experimental tasks and disease correlates, aptly named the Cognitive Atlas. These authors hand-crafted hundreds of terms and their relations with each other, and invited researchers to contribute to documenting new relations – like a Wikipedia for cognitive science. While quality-controlled, curating these processes by hand relies on massive participation of the community, and must match the speed at which new evidence linking old processes is published. A complementary approach to human-generated relations is to let experts judge the validity of machine-generated relations, which can cover much more ground very quickly, saving human time and resources.

Automated Generation of Cognitive Ontology

Here, we present an automated text-mining algorithm that scans through relevant literature databases and builds an

ontology through co-occurrences of cognitive terms mined from the Cognitive Atlas. In particular, we apply the mining algorithm to PubMed, as well as the proceedings to the Annual Meeting of the Cognitive Science Society, in an attempt to automatically generate an ontological structure supplementing the Cognitive Atlas. Furthermore, we search PubMed for cognitive terms in conjunction with neuroimaging terms to establish the cognitive ontology viewed through neuroscience. We note here that previous neuroinformatic works have tackled related challenges. In particular, Yarkoni et al. (2011) created Neurosynth as a meta-analysis of fMRI studies. Its strength lies in providing voxel-level identification of the neural support of cognition, though it necessarily ignores the massive body of electrophysiological research (EEG, MEG, etc.) in favor of certainty in spatial location. In addition, Voytek & Voytek (2012) built BrainSCANR, an automated PubMed text-mining application for similar purposes. However, that work focused primarily on aspects of neuroscience, with inclusion of brain regions and neurochemicals as keywords, while having a limited set of cognitive terms.

In the following sections, we describe the text-mining procedure, as well as an analysis of the word-relations constructed from the automatically generated databases. We present similarities of term-frequency in 4 databases: CogSci (CS), PubMed Cognitive (PMCog), and PubMed Neuro (PMNeu & PMNeuMeth). We further explore latent structures within each database via hierarchical clustering to automatically generate an ontology of cognitive processes.

Methods

All code available online at:
https://github.com/voytekresearch/IdentityCrisis

Data Collection

Term Collection 803 cognitive terms were scraped from the "Concepts" page from the Cognitive Atlas. These were used as the main search terms below, and will thus be referred to as "cognitive terms."

CogSci Abstracts This database is constructed from the title and abstracts of the Presentations, Tutorials, Symposia, and Papers of the Annual Meeting of the Cognitive Science Society from 2010 to 2016. We look for the cognitive terms in each document, constructing a term-document matrix. We then built a co-occurrence matrix by noting all pair-wise co-occurrences of cognitive terms in each document. Data from all 7 years are aggregated. Terms with 50 or more occurrences are included in the clustering analysis (86).

PubMed Cognitive This database is constructed by searching in PubMed for pairs of cognitive terms in quotations, such as "attention"AND"working memory", plus a base phrase: ('AND("cognitive"OR"cognition")'), to ensure searches are constrained to hits relevant to cognition. Counts are recorded as the number of articles that include the search terms in the title or abstracts. Prior to pairs search, we built a term-frequency vector measuring the occurrence of all 803 cognitive terms. Only individual terms with 500 or more hits (217 terms) were included in the pairs search to decrease search time. The number of hits for each pairs of terms (i & j) are recorded in the co-occurrence matrix as element a_{ij}. Search code was built upon the PubMed EUtils Tool API.

PubMed Neuro Method & Neuro These databases are created as the one above, but in conjunction with a base phrase reflecting neuroimaging techniques,
('AND('+ '("fmri"OR"neuroimaging")OR'+
'("pet"OR"positron emission tomography")OR'+
'("eeg"OR"electroencephalography")OR'+
'("meg"OR"magnetoencephalography")OR'+
'("ecog"OR"electrocorticography")OR'+
'("lfp"OR"local field potential")OR'+
'("erp"OR"event related potential")OR'+
'("single unit"OR"single-unit"OR"single neuron")OR'+
'("calcium imaging")')').
138 terms remained after thresholding at 500 hits.
As suggested by reviewers, we further included a "general neuroscience" database that was not exclusively techniques, using ("neural"OR"neuroscience") as base phrase.

Data Analysis

Term-Frequency Term-frequency for each cognitive term were calculated as a fraction by dividing the number of hits a term generated by the total results returned for the base phrase alone (for PubMed databases) or the total number of abstracts (for CogSci database). To visualize differences in term usage, we take the term-frequency difference between pairs of databases and find the terms with the highest absolute difference.

Hierarchical Clustering We use the SciPy hierarchical clustering module (scipy.cluster.hierarchy) to cluster terms based on their normalized co-occurrence matrix, where each row is divided by the diagonal of that row (co-occurrence with self). Dendrograms are generated and leaves are cut (colored) to generate ~N/4 clusters, where N is the total number of terms in tree.

Results

In summary, we find that:
1) there are discrepancies between prevalent terms discovered in the CogSci database and the PubMed Neuro database, with the former leaning towards more theoretical constructs, and the latter, experimentally tangible;
2) hierarchical clustering reveals reasonable yet novel groupings of cognitive terms that are undocumented in the Cognitive Atlas.

Term-Frequency Across Databases

First, we address the question: do cognitive scientists and neuroscientists study the same underlying processes? Table

1 presents the top 20 most frequent cognitive terms in each database.

Table 1: Proportion of term occurrence for the top 20 terms in each database. Green boxes denote terms unique to that database, while red boxes denote terms unique to Neuro.

CogSci		PM Cog		PM Neuro		PM Neuro Meth	
learning	0.257	memory	0.201	activation	0.114	activation	0.11
search	0.228	risk	0.131	learning	0.071	detection	0.068
action	0.218	attention	0.108	memory	0.065	memory	0.06
language	0.185	learning	0.104	loss	0.054	risk	0.059
logic	0.151	association	0.081	action	0.053	attention	0.052
knowledge	0.128	anxiety	0.066	inhibition	0.052	monitoring	0.048
concept	0.126	stress	0.059	attention	0.044	sleep	0.043
memory	0.122	loss	0.054	risk	0.041	association	0.041
context	0.116	activation	0.053	perception	0.038	movement	0.036
decision	0.098	knowledge	0.052	detection	0.037	loss	0.034
attention	0.091	language	0.051	knowledge	0.036	action	0.031
reasoning	0.081	context	0.046	association	0.034	learning	0.03
judgment	0.081	working memory	0.044	context	0.034	perception	0.029
focus	0.079	focus	0.04	stress	0.033	inhibition	0.029
lying	0.074	perception	0.04	movement	0.032	knowledge	0.028
inference	0.072	recognition	0.039	recognition	0.032	focus	0.027
perception	0.072	mood	0.038	induction	0.03	localization	0.027
meaning	0.071	action	0.034	integration	0.029	context	0.027
movement	0.06	pain	0.033	encoding	0.028	recognition	0.026
goal	0.059	sleep	0.032	thought	0.026	language	0.025

First, we note the general trend that CogSci proceedings are much more likely to contain one of several popular terms, with 3 terms appearing in more than 20% of the abstracts and 9 terms in more than 10%. In contrast, only one word in the PMNeuro database is contained in more than 10% of the abstracts ("activation"), which may be artificially inflated due to usages of the word in contexts not describing cognitive activation (e.g., fMRI activation). This suggests that the terms we deem to describe "cognitive processes" do indeed see more usage in the cognitive science community.

On an individual term level, several striking patterns prevail. First, "learning" appears in about 25% of CogSci abstracts, but only 10% in PMCog, 7% in PMNeu, and 3% in PMNeuMeth. This reveals that the concept of "learning" is a rather popular theoretical construct, while being harder to study empirically via neuroimaging. Additionally, "search", "language", and "logic" all appear in more than 15% of CogSci abstracts, but do not crack the 5% mark in PMNeuro, further suggesting the difficulty or reluctance in studying these theoretically important but empirically ill-defined concepts in a neuroscientific context.

On the other hand, "attention", "perception", and "movement" occur in all 4 databases with relatively low but similar proportions. This is unsurprising, as physical processes are much more easily studied in neuroscientific experiments.

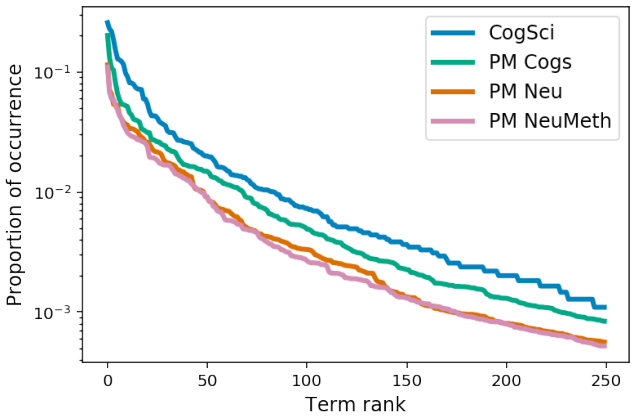

Figure 1: Term frequency results for each database. Note that y-axis is in log scale.

We saw from Table 1 that term usage distribution for the most frequent terms are not the same across the 4 databases. Figure 1 plots these distributions for the top 250 terms used. We see that CogSci proceedings are not very diverse in terms of their topics of investigation, as the more common terms are much more represented in the abstracts. This may be due to the small number of CogSci abstracts available, compared to around 100 times more results returned from PubMed searches. However, PMCogs is less drastic but follows a similar trend, suggesting that cognitive science as a whole refers to these cognitive terms much more frequently than neuroscience.

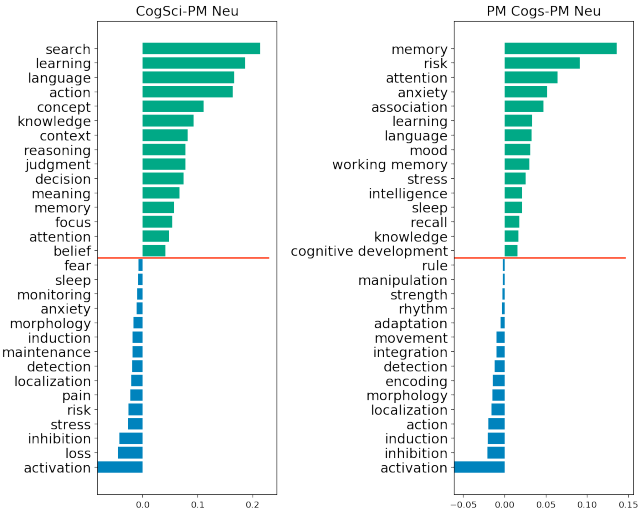

Figure 2. Terms used most differently between Cognitive Science (CogSci, left, PMCogs, right) and Neuroscience.

Finally, we find terms with the biggest usage proportion difference between cognitive science and neuroscience. These results recapitulate the top terms we see in Table 1, where an overwhelming proportion of high-level, conceptual terms are overrepresented in the 2 cognitive datasets. Overall, these analyses demonstrates that, while cognitive terms are adopted more frequently in CogSci

abstracts than the general body of literature in PubMed, many of the processes that are focused on in the CogSci community has not seen as much empirical investigation in the neuroscience community.

Hierarchical Clustering

Having shown a difference in the frequency of cognitive term usage between cognitive science and neuroscience, we turn to the term co-occurrence data to generate ontologies. Here, we can address the question of, in addition to being used with differing frequencies, whether the terms are used in different ways in relation to each other, which suggests a difference in term "meaning". Figure 3 shows dendrograms generated from the CogSci database and PMNeuMeth database. The length of the colored lines (starting from the right) when they merge reflect the similiarity of the the merged terms: the shorter the lines, the more similar they are. As such, pairs of terms like "acuity" and "visual acuity", or "memory" and "working memory" are merged very early on due to the overlap in words, which is a limitation of the text-mining method employed.

Barring these overlapping terms, very reasonable clusters emerge at the mid-level. For example, at the lower end of the CogSci tree exists a language group (red & green) and a learning group (teal). Interestingly, "learning" and "generalization" are very closely tied. Moving up a few clusters, a reasoning cluster emerges (black), including "reasoning", "inference", "induction", and "rule". Similar clusters existing in the PMNeu tree, where the top clusters reflect all forms of perception, then attention, transitioning to speech processing, and finally to language understanding. "Theory of mind" is grouped with "empathy" and "social cognition", while "discrimination" is grouped with "categorization" and "judgement".

Due to the difference in term prevalence between these two databases, some clusters exclusive to one or the other appear. "Logic", "analogy", and "schema" exist as one cluster in the CogSci database, while "anxiety", "fear" and "extinction" emerge as a cluster in the PMNeu database. These clusters clearly reflect the theoretical vs. experimental nature of works published in these two fields. Furthermore, "learning" in CogSci, as mentioned above, talks about a high-level, mental process (tied to "category learning"), while it is linked to "skill", "navigation", and "expertise" in neuroscience. Overall, these examples qualitatively demonstrate that an automated mining and clustering process can tease out: 1) similarity of cognitive terms by grouping them within clusters, and 2) contextualized meaning of terms by grouping them into different clusters specific to cognitive science or neuroscience.

Finally, in keeping with our original goal, we examine whether clusters discovered with our automated process can be used to supplement information in the Cognitive Atlas. Figure 4 demonstrates one example concept: "learning". We observe that the only populated relationship is "are kinds of", in which more specific types of learning are listed. However, the ontological mapping does not capture categorically similar terms described above, such as "generalization" or "categorization". Other examples of missed relationships are more nuanced. For example, under "addiction", the Cognitive Atlas currently includes "reward processing" as a part of addiction (also discovered in our clustering). However, it does not mention "anticipation" and "impulsivity", both of which are key factors in the continuation of addictive behavior. Hence, we conclude that automated clustering of related concepts can greatly aid in the curation of an extensive cognitive ontology.

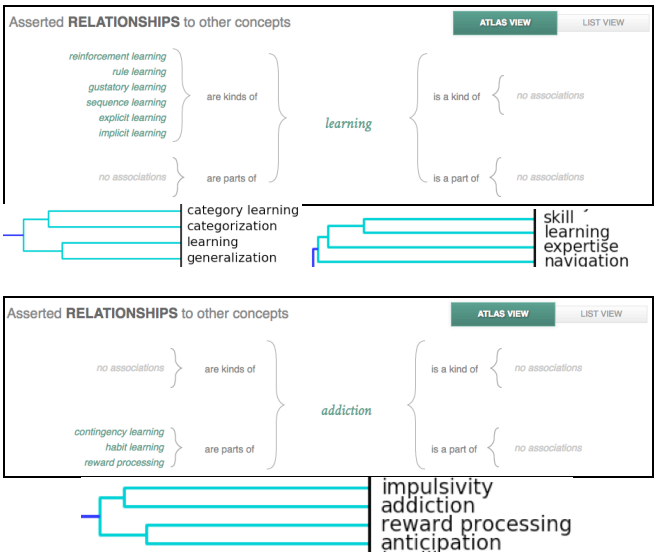

Figure 4: "learning" and "addiction" as curated by Cognitive Atlas, supplemented by clusters generated automatically (from Fig. 3).

Discussion

Summary

In this study, we created a text-mining and clustering pipeline that aims to automate the process of aggregating information from existing literature to create an ontological structure for cognitive processes. We searched for cognitive keywords curated by the Cognitive Atlas, and analyzed databases created by scraping the proceedings to the Annual Cognitive Science meeting, as well as PubMed articles, containing these keywords. We find a prevalent usage of these terms in all the databases, particularly so in the CogSci abstracts. The frequency of term usages differ between CogSci abstracts and PubMed neuroscience articles, likely reflecting the methodological preferences in each field. Hierarchical clustering on pairwise term co-occurrence data group terms relating to each other, demonstrating practicality in serving as a hypothesis-generating procedure to further populate manually-maintained ontologies, such as the Cognitive Atlas.

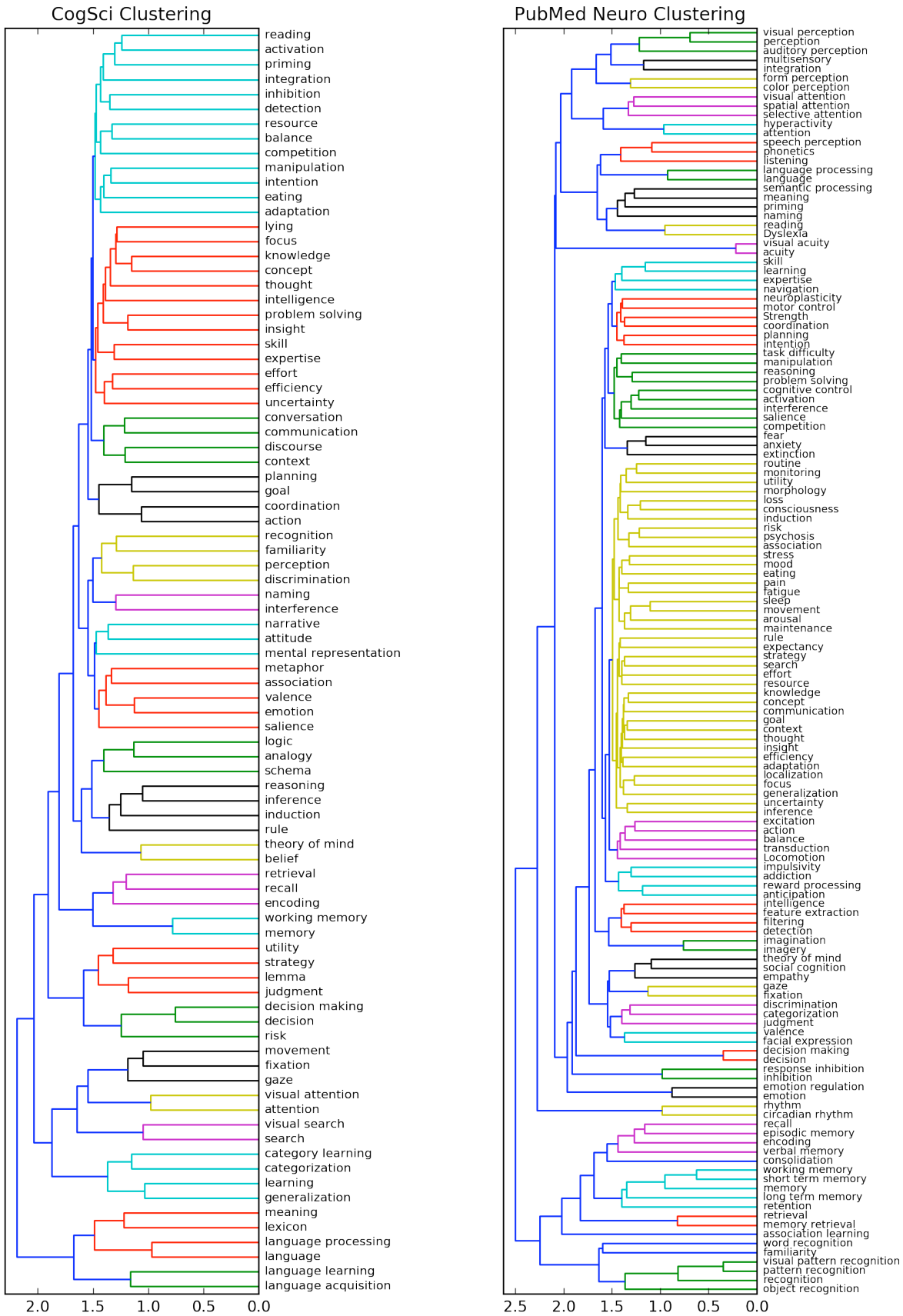

Figure 3: hierarchical clustering results for CogSci and PubMed Neuro Method database.

Implications for Cognitive Science

The current work presents two main contributions. First, the tool itself is completely open-sourced and depends on publicly available databases. Domain-specific researchers can utilize this tool to find common associations to their process of choice, such as addiction. This will be especially useful for beginner researchers, like undergraduate and early graduate students, to quickly situate their topic in the broader context. Furthermore, on a larger scale, this tool can serve as a complementary approach to hand-curated ontologies, saving experts time from manually filling in blanks. One point worth noting is that our work does not attempt to build the ontological structure of cognitive processes as it exists in our minds, similar to ideas suggested by Newell's universal theory of cognition (UTC). Rather, it is a meta-analysis of how cognitive scientists decide to investigate the latent structure of our cognitive processes through their work, with no claims on whether or how this ontological structure actually exists.

Second, the theoretical contribution of this work is that it points to the discrepancy between how cognitive science and neuroscience study cognition. One simple explanation is that neuroscience only partially overlaps with cognitive science, as genetic, molecular, and cellular investigations often do not relate to cognitive phenomena. This is clearly true, however, given that the PubMed Neuro Method database is built specifically with keywords relating to animal neuroimaging techniques, this is unlikely to be the explanation here. Furthermore, the gap similarly exists between PubMed Cognitive and PubMed Neuro databases, so it is not simply a difference in the source of data. Thus, this gap raises the alarming possibility, as one reviewer pointed out, that theories in cognitive science are not testable in the realm of neuroscience, and/or that neuroscience is simply not interested in or ready for the grand theories of cognition.

Limitations & Future Work

While the algorithm returns reasonable and novel results, a few methodological and data-collection limitations must be raised. First, in building the databases, CogSci abstract were collected only up to the annual meeting in 2010, as further archives were unavailable. In contrast, PubMed searches return all hits dating back 30 or more years. As such, it is possible that trends observed in the term-frequency analysis may be due to a temporary peak in interest in certain areas of research, such as "learning". This can be easily ameliorated, however, by rebuilding the PubMed databases while constraining the included search years. In fact, we can analyze different decades (or other periods of time) to see how ontological structure develops over time.

Second, due to the scraping method applied, terms with overlapping words, such as "memory" and "procedural memory" will co-occur with much higher frequency, possibly leading to inflated inferred relationships. Since terms with overlapping words are very likely to have a superset-subset relationship, the over-interpretation of relationship is unlikely to create false positives. However, the artificial increase in co-occurrence may lead clustering to exclude related but now suppressed terms, leading to false negatives. This may be circumvented by making queries for specific terms, i.e., accessing specific rows in the co-occurrence matrix, and ranking related words in their rate of co-occurrence. Hierarchical clustering is simply one method to visualize the co-occurrence data, and many others may be applied on the same dataset to further tease out latent structures, such as Multi-Dimensional Scaling.

Lastly, the co-occurrence matrix is built on the assumption of a bag-of-words model, i.e., word-order and semantic relations don't matter, simply their shared presence in a document. This may lead to spurious linkages, if a document contained a phrase like "attention is not a type of memory." This is likely to be rare, and ultimately, still useful knowledge, as it implies that at some point these terms were wrongfully linked. This last point, however, raises a larger, philosophical question: can automated text mining of existing literature get at the ontology of cognitive science, and if so, is that the same ontology that exists in our minds? We may never know the answer to the latter, but the former is certainly an issue worth investigating. Regardless of whether or not the structure can be recovered from the model presented here, the knowledge structure clearly exists within the minds of practicing cognitive scientists. As such, we may leverage other sources of information, such as citation links, to trace out the ontology, which ultimately just represents a consolidation of knowledge across the broad, interdisciplinary study of cognition.

Acknowledgments

We would like to thank Arturs Semenuks and Torben Noto for creative inspiration, reviewers for their helpful comments and suggestions, as well as the Alfred Sloan Foundation and the Natural Sciences and Engineering Research Council of Canada for funding.

References

Poldrack, R. A. & Yarkoni, T. (2016) From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. *Annul Rev Psychol* **67,** 587–612 (2016).

Poldrack, R. A. *et al.* (2011) The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience. *Front. Neuroinform*.

Voytek, J. B. & Voytek, B. (2012) Automated cognome construction and semi-automated hypothesis generation. *Journal of Neuroscience Methods* **208,** 92–100.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. (2011) Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods* **8,** 665–670.